

Matrix completion by singular value thresholding: sharp bounds

Olga Klopp

CREST and MODAL'X, University Paris Ouest

February 3, 2015

Abstract

We consider the matrix completion problem where the aim is to estimate a large data matrix for which only a relatively small random subset of its entries is observed. Quite popular approaches to matrix completion problem are iterative thresholding methods. In spite of their empirical success, the theoretical guarantees of such iterative thresholding methods are poorly understood. The goal of this paper is to provide strong theoretical guarantees, similar to those obtained for nuclear-norm penalization methods and one step thresholding methods, for an iterative thresholding algorithm which is a modification of the softImpute algorithm. An important consequence of our result is the exact minimax optimal rates of convergence for matrix completion problem which were known until now only up to a logarithmic factor.

Keywords: matrix completion, low rank matrix estimation, minimax optimality

AMS 2000 subject classification: 62J99, 62H12, 60B20, 15A83

1 Introduction

Suppose that we observe a small subset of entries of a large data matrix. The problem of inferring the many missing entries from this small set of observations is known as the matrix completion problem. This problem has attracted considerable attention in the past five years. The first works [7, 6, 5, 11, 21] introduce nuclear-norm minimization method. A different approach, called OPTISPACE has been proposed in [12, 13]. More recently, a method based on max-norm minimization was studied in [4, 10]. Other methods include, for example, GROUSE (Grassmannian Rank-One Update Subspace Estimation) [1] and orthogonal rank-one matrix pursuit [25].

A quite popular direction in the matrix completion literature are the thresholding methods which can be divided in two groups: one-step thresholding methods and iterative thresholding methods. Strong theoretical guarantees were obtained for one-step thresholding procedures. For example, Koltchinskii et al in [17] introduce a soft-thresholding method and show that it is minimax optimal up to a logarithmic factor. In [14] Klopp consider a hard thresholding procedure. Chatterjee [8] propose an universal singular value thresholding that can be applied to a large number of matrix estimation problems, including matrix completion. Despite strong theoretical guarantees these one-step thresholding methods has two important drawbacks: they show poor behavior in practice and only work under the uniform sampling distribution which is not realistic in many practical situations.

Much better practical performances have been shown by iterative thresholding methods. For example, in [3], Cai et al propose a first-order singular value thresholding algorithm SVT which approximately solves the nuclear norm minimization problem. In [19], Mazumder et al introduce **softImpute** algorithm. **softImpute** produces a sequence of solutions that converges to a solution of the nuclear norm regularized least-squares problem when the number of iterations goes to infinity. These iterative thresholding algorithms are simple to implement, scale to relatively large matrices and in practice achieve competitive errors compared to the state-of-the-art algorithms. More recently Dhanjal et al [9] propose an improvement for the **softImpute** algorithm using randomized SVDs along with a novel updating method. This improvement allows to bypass the bottleneck in the algorithm which consists in the use of the singular value decomposition of a large matrix at each iteration.

The majority of existing algorithms for matrix completion are batch methods, that is, they operate on the full data matrix. However in some applications such as recommendation systems or localization in sensor networks we observe a sequence of data matrix M_1, \dots, M_T revealed sequentially where from M_t to M_{t+1} we add new observations. In such situations the predictive rule should be refined incrementally. One advantage of iterative thresholding algorithms is that they can be adapted to such sequential learning, see for example [9].

In spite of their empirical success, the theoretical guarantees of such iterative thresholding methods are poorly understood. The goal of this paper is to provide strong theoretical guarantees, similar to those obtained for nuclear-norm penalization methods (see, for example [20, 15]) and one step thresholding methods (see [17, 14, 8]) for a modification of the **softImpute** algorithm.

1.1 Contributions and Related Work

The contributions of the present paper to the theoretical study of the modified **softImpute** algorithm are multifaceted. In Section 3.2 we prove an upper bound on the estimation error of the output \hat{M} of our algorithm. Let $M_0 \in \mathbb{R}^{m_1 \times m_2}$ be the unknown matrix of interest. Suppose, for simplicity, that each entry is observed with the same probability p , then we prove the following upper bound

on the estimation error of \hat{M}

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \lesssim \frac{\text{rank}(M_0)}{p \min(m_1, m_2)}. \quad (1)$$

Here the symbol \lesssim means that the inequality holds up to a multiplicative numerical constant. To the best of our knowledge, the upper bound on the estimation error given by (1) is strictly better than all upper bounds available in matrix completion literature.

For instance, for the same setting, Chatterjee in [8] obtains the following larger bound

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \lesssim \sqrt{\frac{\text{rank}(M_0)}{p \min(m_1, m_2)}}.$$

On the other hand, [17, 20, 15], among some other papers, consider a slightly different setting where the matrix completion problem is viewed as a particular case of the trace regression model. In this setting the number of observations n is fixed. The drawback here is that in this model each entry can be observed multiple times which is not the case in a large number of practical situations. We consider a different setting where each entry can be observed at most once (see Section 2.1). However, it is easy to see that these two settings are closely related if we put $n = pm_1 m_2$. Comparing to (1), the bounds obtained in [17, 20, 15] have an additional $\log(d_1 + d_2)$ factor.

Koltchinskii et al in [17] obtained lower bounds for the estimation error without this additional $\log(d_1 + d_2)$ factor. So our result answer the important theoretical question what is the exact minimax rate of convergence for matrix completion problem. As the lower bound in [17] is obtained for a different setting, in Section 4 we adapt their proof to our setting, showing that the minimax rate of convergence for matrix completion problem is given by (1) and that the estimator produced by our algorithm is minimax optimal. Note that our techniques can be adapted to the setting considered in [17, 20, 15] and lead to an upper bound without the additional $\log(d_1 + d_2)$ factor in this setting also.

Another important point is that a large part of matrix completion literature consider uniform sampling at random setting where each entry is observed with the same probability p . In many applications, such as recommendation systems, this assumption is not realistic. The theoretical analysis in the present paper is carried out for quite general sampling distributions and show that our iterative thresholding algorithm has good performances in such situations. Finally our results give theoretical insights for the chose of the parameters in the modified `softImpute` algorithm.

1.2 Organisation of the paper

The remainder of this paper is organized as follows. In Section 2.1 we introduce our model and the assumptions on the sampling scheme. For the reader's convenience, we collect notation which we use throughout the paper in Section 2.2.

In Section 3.1 we present a modification of the `softImpute` algorithm for matrix completion. The upper bounds on the estimation error are derived in Section 3.2. Finally the lower bounds are obtained in Section 4 and the Appendix contains the proofs.

2 Preliminaries

2.1 Model and sampling scheme

Suppose that we observe a relatively small number of entries of a data matrix

$$X = M_0 + E. \quad (2)$$

Here $M_0 = (m_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is the unknown matrix of interest and $E = (\xi_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is the matrix containing the noise. We assume that the noise variables ξ_{ij} are independent, zero mean and bounded:

Assumption 1. $\mathbb{E}(\xi_{ij}) = 0$, $\mathbb{E}(\xi_{ij}^2) = \sigma^2$ and there exists a positive constant $b > 0$ such that

$$\max_{i,j} |\xi_{ij}| \leq b.$$

We suppose that each entry of X is observed independently of the other entries. For the entry $(i, j) \in [m_1] \times [m_2]$, we denote the probability to be observed by π_{ij} . Let η_{ij} be the independent Bernoulli variables with parameters π_{ij} and $y_{ij} = \eta_{ij}(m_{ij} + \xi_{ij})$. Then, $Y = (y_{ij})$ is the matrix containing our observations. We denote by Ω the random set of observed indices.

In the simplest situation each coefficient is observed with the same probability, i.e. for every $(i, j) \in [m_1] \times [m_2]$, $\pi_{ij} = p$. Unfortunately, such an assumption on the sampling distribution is not realistic in many practical applications. In the present paper, we consider general sampling model. We suppose that each coefficient is observed with a positive probability:

Assumption 2. There exists $p > 0$ such that for any $(i, j) \in \{1, \dots, m_1\} \times \{1, \dots, m_2\}$

$$\pi_{ij} \geq p.$$

For any $A = (A_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ we define the weighted by π_{ij} Frobenius norm of A

$$\|A\|_{L_2(\Pi)}^2 = \sum_{(i,j)} \pi_{ij} A_{ij}^2.$$

Assumption 2 implies that

$$\|A\|_{L_2(\Pi)}^2 \geq p^{-1} \|A\|_2^2. \quad (3)$$

We denote the column and row marginals by

$$\pi_{\cdot j} = \sum_{i=1}^{m_1} \pi_{ij} \quad \text{and} \quad \pi_{i \cdot} = \sum_{j=1}^{m_2} \pi_{ij}.$$

Suppose that we know an upper bound L on it's maximum:

$$\max_{i,j} (\pi_{.j}, \pi_{i.}) \leq L. \quad (4)$$

Note that we can easily get an estimation on this upper bound using the empirical frequencies

$$\hat{\pi}_{.j} = \frac{\sum_{i=1}^{m_1} \eta_{ij}}{\sum_{(i,j)} \eta_{ij}} \quad \text{and} \quad \hat{\pi}_{i.} = \frac{\sum_{j=1}^{m_2} \eta_{ij}}{\sum_{(i,j)} \eta_{ij}}.$$

2.2 Notation

We provide a brief summary of the notation used throughout this paper. Let A, B be matrices in $\mathbb{R}^{m_1 \times m_2}$.

- For a matrix A , A_{ij} is its (i, j) th entry.
- We denote by $S_\lambda(W) \equiv UD_\lambda V'$ the *soft-thresholding* operator where $D_\lambda = \text{diag}[(d_1 - \lambda)_+, \dots, (d_r - \lambda)_+]$, UDV' is the SVD of W , $D = \text{diag}[d_1, \dots, d_r]$ and $t_+ = \max(t, 0)$.
- For any set I , $|I|$ denotes its cardinal and \bar{I} its complement. Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.
- For two matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$ we define the *scalar product*

$$\langle A, B \rangle = \text{tr}(A^T B).$$

- We denote by $\|A\|_2$ the usual l_2 -norm. Additionally, we use the following matrix norms: $\|A\|_*$ is the nuclear norm (the sum of singular values), $\|A\|$ is the operator norm (the largest singular value), $\|A\|_\infty$ is the largest absolute value of the entries:

$$\|A\|_\infty = \max_{i,j} |A_{ij}|.$$

- π_{ij} is the probability to observe the (i, j) -th element. For $j = 1 \dots m_2$, $\pi_{.j} = \sum_{i=1}^{m_1} \pi_{ij}$ and for $i = 1 \dots m_1$, $\pi_{i.} = \sum_{j=1}^{m_2} \pi_{ij}$. We have that

$$\max_{i,j} (\pi_{.j}, \pi_{i.}) \leq L.$$

- Let $M = \max(m_1, m_2)$, $m = \min(m_1, m_2)$ and $d = m_1 + m_2$.
- Let $I \subset \{1, \dots, m_1\} \times \{1, \dots, m_2\}$ be a subset of indices. Given a matrix $A = (A_{ij})$, we define its restriction on I , A_I , in the following way: $(A_I)_{ij} = A_{ij}$ if $(ij) \in I$ and $(A_I)_{ij} = 0$ if not.

- We denote $\|A\|_{L_2(\Pi)}^2 = \sum_{(i,j)} \pi_{ij} A_{ij}^2$ and Assumption 2 implies

$$\|A\|_{L_2(\Pi)}^2 \geq p^{-1} \|A\|_2^2.$$

- Let $\{\epsilon_{ij}\}$ be an i.i.d. Rademacher sequence and $X_{ij} = e_i(m_1) e_j^*(m_2)$ where $e_k(l)$ are the canonical basis vectors in \mathbb{R}^l . We define

$$\Sigma_R = \sum_{(i,j)} \eta_{ij} \epsilon_{ij} X_{ij} \quad \text{and} \quad \Sigma = \sum_{(i,j)} \eta_{ij} \xi_{ij} X_{ij}. \quad (5)$$

3 The Singular Value Thresholding Algorithm

In this section we introduce an iterative singular value thresholding algorithm and discuss its theoretical properties. We show that it enjoys strong theoretical guarantees and, unlike one-step thresholding procedures, is well adapted for general non-uniform sampling distributions.

3.1 Algorithm

Our algorithm is based on the **softImpute** algorithm proposed by Mazumder et al in [19]. **SoftImpute** algorithm is inspired by SVD-Impute of Troyanskaya et al [23]. It alternates between imputing the missing values from a current SVD, and updating the SVD using the data matrix.

Algorithm 1

Require : Matrix Y , regularization parameter λ and a , an upper bound on the sup-norm of M_0 .

1. $M^{old} = 0$
2. (a) Repeat
 - (i) Compute $M^{new} \leftarrow S_\lambda(Y + (M^{old})_\Omega)$.
 - (ii) If $\|(M^{new} - M^{old})_{\bar{\Omega}}\| < \lambda/3$ and $\|M^{new} - M^{old}\|_\infty < a$ exit.
 - (iii) Put $M^{old} = (M_{ij}^{old})$

$$M_{ij}^{old} = \begin{cases} M_{ij}^{new} & \text{if } |M_{ij}^{new}| \leq a \\ a & \text{if } M_{ij}^{new} > a \\ -a & \text{if } M_{ij}^{new} < -a. \end{cases} \quad (6)$$

- (b) Assign $\hat{M} \leftarrow M^{new}$.
3. Output \hat{M} .

This algorithm repeatedly replaces the missing entries with the current guess, update the guess by solving

$$M^{new} \in \underset{M}{\text{minimize}} f_\lambda(M) = \frac{1}{2} \|Y + (M^{old})_{\bar{\Omega}} - M\|_2^2 + \lambda \|M\|_* \quad (7)$$

and truncating M^{new} . Let us denote by $(M_k)_{k \geq 0}$ the sequence of solutions produced by Algorithm 1. We have the following result :

Lemma 1. *For the successive differences of the sequence $(M_k)_{k \geq 0}$ we have that*

$$\|M^{k+1} - M^k\|_2 \rightarrow 0 \text{ as } k \rightarrow \infty \quad (8)$$

which implies

$$\|(M^{k+1} - M^k)_{\bar{\Omega}}\| \rightarrow 0 \quad \text{and} \quad \|M^{k+1} - M^k\|_\infty \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (9)$$

3.2 Upper bound on the estimation error

In this section we derive an upper bound on the estimation error of \hat{M} produced by Algorithm 1. This bound is non-asymptotic and implies, in particular, that the proposed estimator is minimax optimal. We start by a general result which is proven in Appendix A.

Theorem 2. *Let Assumptions 1 and 2 be satisfied and $\|M_0\|_\infty \leq a$. Assume that $\lambda \geq 3 \|\Sigma\|$. Then, with probability at least $1 - 8/d$,*

$$\|\hat{M} - M_0\|_{L_2(\Pi)}^2 \leq C p^{-1} \left\{ \text{rank}(M_0) \left(\lambda^2 + a^2 (\mathbb{E}(\|\Sigma_R\|))^2 \right) + a^2 + \log(d) \right\}.$$

where $d = m_1 + m_2$.

Using Assumption 2, Theorem 2 implies the following bound on the estimation error measured in normalized Frobenius norm

Corollary 3. *Under assumptions of Theorem 2 and with probability at least $1 - 8/d$,*

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \leq \frac{C}{p^2 m_1 m_2} \left\{ \text{rank}(M_0) \left(\lambda^2 + a^2 (\mathbb{E}(\|\Sigma_R\|))^2 \right) + a^2 + \log(d) \right\}.$$

In order to get a bound in a closed form we need to obtain a suitable upper bounds on $\mathbb{E}(\|\Sigma_R\|)$ and, with probability close to 1, on $\|\Sigma\|$.

Lemma 4. *Suppose that (ξ_{ij}) are independent and satisfy Assumption 1. Then, there exists absolute constants $c^*, C^* > 0$ such that, for all $t > 0$ with probability at least $1 - me^{-t^2}$ we have*

$$\|\Sigma\| \leq 3\sigma\sqrt{2L} + c^* b t \quad (10)$$

where $L \leq 1$ is defined in (4).

Moreover, we have

$$\mathbb{E} \|\Sigma_R\| \leq C^* \left(\sqrt{L} + \sqrt{\log m} \right). \quad (11)$$

This Lemma is proven in Appendix F.

Taking $t = \sqrt{2 \log(d)}$ in Lemma 4, we get that with probability at least $1 - 1/d$,

$$\|\Sigma\| \leq 3\sigma\sqrt{2L} + c^*b\sqrt{2 \log(d)},$$

then, we can choose

$$\lambda = 3 \left(3\sigma\sqrt{2L} + c^*b\sqrt{2 \log(d)} \right). \quad (12)$$

With this choice of λ we obtain the following Theorem.

Theorem 5. *Let Assumptions 1 and 2 be satisfied and $\|M_0\|_\infty \leq a$. Then, with probability at least $1 - 8/d$,*

$$\|\hat{M} - M_0\|_{L_2(\Pi)}^2 \leq C p^{-1} \text{rank}(M_0) \left\{ (a \vee \sigma)^2 L + a^2 \log(m) + b^2 \log(d) \right\}.$$

and

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \leq \frac{C \text{rank}(M_0)}{p^2 m_1 m_2} \left\{ (a \vee \sigma)^2 L + a^2 \log(m) + b^2 \log(d) \right\}.$$

Remark 1. Note that $\pi_{ij} \geq p$ yields $L \geq Mp$. Then, the upper bound on the estimation error in the Theorem 5 is at least a constant times $\frac{\text{rank}(M_0)}{pm}$.

So, in order to get a small estimation error, p should be larger than $\frac{\text{rank}(M_0)}{m}$.

We denote by $n = \sum_{ij} \pi_{ij}$ the expected number of observations. Condition $p \geq \frac{\text{rank}(M_0)}{m}$ implies the following condition on n

$$n \geq C \text{rank}(M_0) M. \quad (13)$$

When the rank of the matrix M_0 is small, this necessary number of observations is close to the number of degree of freedom of the matrix M_0 , which is

$$(m_1 + m_2) \text{rank}(M_0) - (\text{rank}(M_0))^2.$$

Let us restrict our attention to the non-degenerated case $M_0 \neq 0$ (we can easily include this case replacing $\text{rank}(M_0)$ by $\text{rank}(M_0) \vee 1$). Assuming that the expected number of observations n is not too small, we can get simpler bound on the estimation error. Suppose that $n > c^* m \log(d)$. Then, using

$$Lm \geq n \geq c^* m \log d$$

we get $L \geq c^* \log d$ and we can chose λ in the following way

$$\lambda = 18b\sqrt{2L}. \quad (14)$$

With this choice of λ we get the following bound on the estimation error

Corollary 6. *Let Assumptions 1 and 2 be satisfied and $\|M_0\|_\infty \leq a$. Assume that $n \geq c^* m \log(d)$ and $M_0 \neq 0$. Then, with probability at least $1 - 8/d$,*

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \leq \frac{C \text{rank}(M_0) (a \vee b)^2 L}{p^2 m_1 m_2}.$$

In order to compare this result with previous results on noisy matrix completion we consider a more restrictive assumption on the sampling distribution. That is, we assume that this distribution is close to the uniform one:

Assumption 3. *There exists positives constants μ_1 and μ_2 independent on m_1 and m_2 and a $0 < p < 1$ such that for every $(i, j) \in \{1, \dots, m_1\} \times \{1, \dots, m_2\}$ we have*

$$\mu_2 p \leq \pi_{ij} \leq \mu_1 p.$$

Under this assumption Theorem 2 yields

Corollary 7. *Let Assumptions 1 and 3 be satisfied and $\|M_0\|_\infty \leq a$. Assume that $n \geq m \log(d)$ and λ given by (14). Then, with probability at least $1 - 8/d$,*

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \leq \frac{C \text{rank}(M_0) (a \vee b)^2}{pm}.$$

Remark 2. Let us compare the bound given by Corollary 7 with bounds available in the literature. Our model was previously considered by Chatterjee in [8] in the case of uniform sampling distribution, that is $\pi_{ij} = p$ for any $(i, j) \in \{1, \dots, m_1\} \times \{1, \dots, m_2\}$. In [8], Chatterjee introduces a simple estimation procedure, called Universal Singular Value Thresholding which is applied to a number of questions in low rank matrix estimation, blockmodels, distance matrix completion, latent space models and etc. For matrix completion problem and under the additional assumption $p \geq n^{-1+\epsilon}$ for some $\epsilon > 0$, the bound obtained in [8] is the following one

$$\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} \leq C \sqrt{\frac{\text{rank}(M_0) (a \vee b)^2}{pm}}.$$

The rate of convergence given by Corollary 7 is faster and, as we will see in Section 4, is minimax optimal. Note that the additional assumption $p \geq n^{-1+\epsilon}$ yields the following condition on the expected number of observations

$$n > m^\epsilon M. \tag{15}$$

For low rank matrices, this necessary number of observations is larger than the number of observations required by our method and given by (13).

In [20, 17, 15] a closely related set up for matrix completion problem using the trace regression model was considered. The main difference between these two settings is that in the case of the trace regression the number of observations

is not random and each entry may be observed multiple times. In our setting the number of observations is random and each entry is observed at most once. Comparing with Corollary 7 and using $n = pm_1m_2$ we see that bounds obtained in [20, 17, 15] contain an additional logarithmic factor $\log(m_1 + m_2)$.

4 Minimax Lower bounds

In this section, we prove the minimax lower bound showing that the rates attained by our estimator are optimal. The minimax lower bound in a closely related problem was obtained by Koltchinskii et al in [17]. We adapt their proof to our set up.

We will denote by $\inf_{\hat{M}}$ the infimum over all the estimators. For any $M_0 \in \mathbb{R}^{m_1 \times m_2}$, let \mathbb{P}_{M_0} denote the probability distribution of the observations

$$(\eta_{11}X_{11}, \dots, \eta_{m_1m_2}X_{m_1m_2})$$

satisfying (2).

For any integer $0 \leq r \leq \min(m_1, m_2)$ and any $a > 0$, we consider the class of matrices

$$\mathcal{A}(r, a) = \{M \in \mathbb{R}^{m_1 \times m_2} : \text{rank}(M) \leq r, \|M\|_\infty \leq a, \}. \quad (16)$$

We will prove the lower bound in the case of the uniform sampling distribution, that is, we suppose that each entry is observed with the same probability p . As it was noted in Remark 1, in order to get a small estimation error we need to observe a sufficiently large number of entries, or, equivalently, the probability p should be larger than r/m . We prove a lower bound on the estimation risk when this condition is satisfied.

Theorem 8. *Suppose that $m_1, m_2 \geq 2$ and $p \geq \frac{r}{m}$. Fix $a > 0$ and integer $1 \leq r \leq \min(m_1, m_2)$. Suppose that the variables ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, for $i = 1, \dots, n$. Then, there exist absolute constants $\beta \in (0, 1)$ and $c > 0$, such that*

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{A}(r, a)} \mathbb{P}_{M_0} \left(\frac{\|\hat{M} - M_0\|_2^2}{m_1m_2} > \frac{cr(a \wedge \sigma)^2}{pm} \right) \geq \beta.$$

A Proof of Theorem 2

1. By Lemma 1 in [19], \hat{M} minimizes

$$f_\lambda(M) = \frac{1}{2} \|Y + (M^{old})_{\hat{\Omega}} - M\|_2^2 + \lambda \|M\|_*.$$

Then, using the sub-gradient stationary conditions we have

$$-\langle Y + (M^{old})_{\hat{\Omega}} - \hat{M}, \hat{M} - M_0 \rangle + \lambda \langle \hat{V}, \hat{M} - M_0 \rangle \leq 0$$

where $\hat{V} \in \partial\|\hat{M}\|_*$. A simple calculation yields

$$\begin{aligned} \left\| (M_0 - \hat{M})_{\Omega} \right\|_2^2 &\leq \underbrace{\left| \left\langle (Y - M_0)_{\Omega}, \hat{M} - M_0 \right\rangle \right|}_{\text{I}} + \underbrace{\left| \left\langle (M^{old} - \hat{M})_{\bar{\Omega}}, \hat{M} - M_0 \right\rangle \right|}_{\text{II}} \\ &\quad + \underbrace{\lambda \left\langle \hat{V}, M_0 - \hat{M} \right\rangle}_{\text{III}}. \end{aligned} \quad (17)$$

2. We estimate each term in (17) separately. For the first term, we have that $(Y - M_0)_{\Omega} = \Sigma$ where $\Sigma = \sum_{(i,j)} \eta_{ij} \xi_{ij} X_{ij}$. Then, by the duality between the nuclear and the operator norms, we obtain

$$\left| \left\langle (Y - M_0)_{\Omega}, \hat{M} - M_0 \right\rangle \right| \leq \|\Sigma\| \|\hat{M} - M_0\|_*. \quad (18)$$

For the second term, using again the duality between the nuclear and the operator norms and the stopping criteria for the Algorithm 1, we obtain

$$\begin{aligned} \left| \left\langle (M^{old} - \hat{M})_{\bar{\Omega}}, \hat{M} - M_0 \right\rangle \right| &\leq \left\| (M^{old} - \hat{M})_{\bar{\Omega}} \right\| \left\| \hat{M} - M_0 \right\|_* \\ &\leq \lambda/3 \left\| \hat{M} - M_0 \right\|_*. \end{aligned} \quad (19)$$

3. In order to estimate the third term, we use that by monotonicity of subdifferentials of convex functions we have that $\left\langle \hat{V} - V, \hat{M} - M_0 \right\rangle \geq 0$, for any $V \in \partial\|M_0\|_*$. This implies

$$\left\langle \hat{V}, M_0 - \hat{M} \right\rangle \leq \left\langle V, M_0 - \hat{M} \right\rangle. \quad (20)$$

Let P_S be the projector on the linear vector subspace S and let S^{\perp} be the orthogonal complement of S . Let $u_j(A)$ and $v_j(A)$ denote respectively the *left* and *right* orthonormal *singular vectors* of a matrix A . $S_1(A)$ is the linear span of $\{u_j(A)\}$, $S_2(A)$ is the linear span of $\{v_j(A)\}$. We set

$$\mathbf{P}_A^{\perp}(B) = P_{S_1^{\perp}(A)} B P_{S_2^{\perp}(A)} \quad \text{and} \quad \mathbf{P}_A(B) = B - \mathbf{P}_A^{\perp}(B). \quad (21)$$

Since $\mathbf{P}_A(B) = P_{S_1^{\perp}(A)} B P_{S_2(A)} + P_{S_1(A)} B$ and $\text{rank}(P_{S_i(A)} B) \leq \text{rank}(A)$ we have that

$$\text{rank}(\mathbf{P}_A(B)) \leq 2 \text{rank}(A). \quad (22)$$

Note that the subdifferential of the convex function $A \rightarrow \|A\|_*$ is the following set of matrices (cf. [26])

$$\partial\|A\|_* = \left\{ \sum_{j=1}^{\text{rank}(A)} u_j(A) v_j^T(A) + \mathbf{P}_A^{\perp}(W) : \|W\| \leq 1 \right\}. \quad (23)$$

Inequality (19) and (23) imply

$$\mathbf{III} \leq \lambda \left\langle \sum_{j=1}^R u_j(M_0) v_j^T(M_0), M_0 - \hat{M} \right\rangle + \left\langle \mathbf{P}_{M_0}^\perp(W), M_0 - \hat{M} \right\rangle. \quad (24)$$

Using the fact that $\left\| \sum_{j=1}^R u_j(M_0) v_j^T(M_0) \right\| = 1$ and

$$\left\langle \sum_{j=1}^R u_j(M_0) v_j^T(M_0), M_0 - \hat{M} \right\rangle = \left\langle \sum_{j=1}^R u_j(M_0) v_j^T(M_0), \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\rangle$$

we obtain

$$\mathbf{III} \leq \lambda \left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_* + \left\langle \mathbf{P}_{M_0}^\perp(W), M_0 - \hat{M} \right\rangle. \quad (25)$$

Now, by the duality between the nuclear and the operator norms, there exists W with $\|W\| \leq 1$ and such that

$$\left\langle \mathbf{P}_{M_0}^\perp(W), M_0 - \hat{M} \right\rangle = - \left\langle W, \mathbf{P}_{M_0}^\perp(\hat{M}) \right\rangle = - \left\| \mathbf{P}_{M_0}^\perp(\hat{M}) \right\|_*. \quad (26)$$

For this particular choice of W , (25) and (26) imply

$$\mathbf{III} \leq \lambda \left(\left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_* - \left\| \mathbf{P}_{M_0}^\perp(\hat{M}) \right\|_* \right). \quad (27)$$

Putting (18), (19), and (27) into (17) and using $\lambda \geq 3 \|\Sigma\|$ we obtain

$$\left\| (M_0 - \hat{M})_\Omega \right\|_2^2 \leq \frac{2\lambda}{3} \left\| \hat{M} - M_0 \right\|_* + \lambda \left(\left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_* - \left\| \mathbf{P}_{M_0}^\perp(\hat{M}) \right\|_* \right). \quad (28)$$

4. The triangle inequality and (22) lead to

$$\left\| (M_0 - \hat{M})_\Omega \right\|_2^2 \leq \frac{5\lambda}{3} \left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_* \leq \frac{5\lambda \sqrt{2\text{rank}(M_0)}}{3} \left\| M_0 - \hat{M} \right\|_2 \quad (29)$$

and

$$\frac{\lambda}{3} \left\| \mathbf{P}_{M_0}^\perp(\hat{M}) \right\|_* \leq \frac{5\lambda}{3} \left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_*. \quad (30)$$

Inequality (30) implies

$$\left\| \mathbf{P}_{M_0}^\perp(\hat{M}) \right\|_* \leq 5 \left\| \mathbf{P}_{M_0}(M_0 - \hat{M}) \right\|_*$$

and

$$\left\| \hat{M} - M_0 \right\|_* \leq 6 \left\| \mathbf{P}_{M_0}(\hat{M} - M_0) \right\|_* \leq \sqrt{72 \text{rank}(M_0)} \left\| \hat{M} - M_0 \right\|_2. \quad (31)$$

5. For a $0 < r \leq m$ we consider the following constrain set

$$\mathcal{C}(r) = \left\{ A \in \mathbb{R}^{m_1 \times m_2} : \|A\|_\infty = 1, \|A\|_{L_2(\Pi)}^2 \geq \frac{\log(d)}{0.0006 \log(6/5) p}, \|A\|_* \leq \sqrt{r} \|A\|_2 \right\}. \quad (32)$$

Note that the condition $\|A\|_* \leq \sqrt{r} \|A\|_2$ is satisfied if $\text{rank}(A) \leq r$.

We have the following result for matrices in $\mathcal{C}(r)$. Its proof is given in Appendix C.

Lemma 9. *For all $A \in \mathcal{C}(r)$*

$$\|A_\Omega\|_2^2 \geq \frac{\|A\|_{L_2(\Pi)}^2}{2} - 44 p^{-1} \left[r (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right]$$

with probability at least $1 - 8/d$.

Note that condition $\left\| \hat{M} - M^{old} \right\|_\infty < a$ and $\|M^{old}\|_\infty < a$ imply

$$\left\| \hat{M} - M_0 \right\|_\infty \leq 3a.$$

We now consider two cases, depending on whether the matrix $\frac{(\hat{M} - M_0)}{3a}$ belongs to the set $\mathcal{C}(72 \text{rank}(M_0))$ or not.

Case 1: Suppose first that $\left\| \hat{M} - M_0 \right\|_{L_2(\Pi)}^2 < \frac{\log(d)}{0.0006 \log(6/5) p}$, then the statement of the Theorem 2 is true.

Case 2: It remains to consider the case $\left\| \hat{M} - M_0 \right\|_{L_2(\Pi)}^2 \geq \frac{\log(d)}{0.0006 \log(6/5) p}$.

Then (31) implies that $\frac{1}{3a} (\hat{M} - M_0) \in \mathcal{C}(72 \text{rank}(M_0))$ and we can apply Lemma 9. From Lemma 9 and (29) we obtain that with probability at least $1 - 8/d$ one has

$$\begin{aligned} \frac{1}{2} \left\| \hat{M} - M_0 \right\|_{L_2(\Pi)}^2 &\leq \frac{5\lambda \sqrt{2 \text{rank}(M_0)}}{3} \left\| M_0 - \hat{M} \right\|_2 \\ &\quad + 369 a^2 p^{-1} \left[72 \text{rank}(M_0) (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right] \\ &\leq 6\lambda^2 p^{-1} \text{rank}(M_0) + \frac{p}{4} \left\| \hat{M} - M_0 \right\|_2^2 \\ &\quad + 369 a^2 p^{-1} \left[72 \text{rank}(M_0) (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right]. \end{aligned}$$

Now (3) imply that, there exist numerical constants C such that

$$\left\| \hat{M} - M_0 \right\|_{L_2(\Pi)}^2 \leq C p^{-1} \left\{ \text{rank}(M_0) \left(\lambda^2 + a^2 (\mathbb{E}(\|\Sigma_R\|))^2 \right) + a^2 \right\},$$

which leads to the statement of the Theorem 2.

B Proof of Theorem 8

We adopt the proof of Theorem 5 in [17] to our setting. Assume w.l.o.g. that $m_1 \geq m_2$. For a $\gamma \leq 1$, define

$$\tilde{\mathcal{L}} = \left\{ \tilde{L} = (l_{ij}) \in \mathbb{R}^{m_1 \times r} : l_{ij} \in \left\{ 0, \gamma(\sigma \wedge a) \left(\frac{r}{pm} \right)^{1/2} \right\}, \forall 1 \leq i \leq m_1, 1 \leq j \leq r \right\},$$

and consider the associated set of block matrices

$$\mathcal{A} = \left\{ L = (\tilde{L} \mid \dots \mid \tilde{L} \mid O) \in \mathbb{R}^{m_1 \times m_2} : \tilde{L} \in \tilde{\mathcal{L}} \right\},$$

where O denotes the $m_1 \times (m_2 - r \lfloor m_2 / (2r) \rfloor)$ zero matrix, and $\lfloor x \rfloor$ is the integer part of x .

Remark 3. *In the case $m_1 < m_2$, we only need to change the construction of the low rank component of the test set. We first build a matrix $\tilde{L} = (\tilde{L} \mid O) \in \mathbb{R}^{r \times m_2}$ where $\tilde{L} \in \mathbb{R}^{r \times (m_2/2)}$ with entries in $\left\{ 0, \gamma(\sigma \wedge a) \left(\frac{r}{pm} \right)^{1/2} \right\}$ and, then, we replicate this matrix to obtain a block matrix L of size $m_1 \times m_2$*

$$L = \begin{pmatrix} \tilde{L} \\ \vdots \\ \tilde{L} \\ O \end{pmatrix}.$$

By construction, any element of \mathcal{A} as well as the difference of any two elements of \mathcal{A} has rank at most r . In addition, condition $p \geq \frac{r}{m}$ implies that the entries of any matrix in \mathcal{A} take values in $[0, a]$. Thus, $\mathcal{A} \subset \mathcal{A}(r, a)$.

The Varshamov-Gilbert bound (cf. Lemma 2.9 in [24]) guarantees the existence of a subset $\mathcal{A}^0 \subset \mathcal{A}$ with cardinality $\text{Card}(\mathcal{A}^0) \geq 2^{(rM)/8} + 1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements A_1 and A_2 of \mathcal{A}^0 ,

$$\|A_1 - A_2\|_2^2 \geq \frac{Mr}{8} \left(\gamma^2(\sigma \wedge a)^2 \frac{r}{pm} \right) \left\lfloor \frac{m}{r} \right\rfloor \geq \frac{\gamma^2}{16} (\sigma \wedge a)^2 m_1 m_2 \frac{r}{pm}. \quad (33)$$

Using that, conditionally on X_i , the distributions of ξ_i are Gaussian, we get that, for any $A \in \mathcal{A}_0$, the Kullback-Leibler divergence $K(\mathbb{P}_0, \mathbb{P}_A)$ between \mathbb{P}_0 and \mathbb{P}_A satisfies

$$K(\mathbb{P}_0, \mathbb{P}_A) = \frac{1}{2\sigma^2} \|A\|_{L_2(\Pi)}^2 \leq \frac{\gamma^2 Mr}{2}. \quad (34)$$

From (34) we deduce that the condition

$$\frac{1}{\text{Card}(\mathcal{A}^0) - 1} \sum_{A \in \mathcal{A}^0} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\text{Card}(\mathcal{A}^0) - 1) \quad (35)$$

is satisfied for any $\alpha > 0$ if $\gamma > 0$ is chosen as a sufficiently small numerical constant depending on α . In view of (33) and (35) and using the application of Theorem 2.5 in [24] implies

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{A}(r, a)} \mathbb{P} \left(\frac{\|\hat{M} - M_0\|_2^2}{m_1 m_2} > \frac{C(\sigma \wedge a)^2 r}{pm} \right) \geq \beta \quad (36)$$

for some absolute constants $\beta \in (0, 1)$, which implies the statement of Theorem 8.

C Proof of Lemma 9

This proof is close to the proof of Lemma 12 in [15]. Set

$$\mathcal{E} = 44 p^{-1} \left[r (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right].$$

We will show that the probability of the following “bad” event is small

$$\mathcal{B} = \left\{ \exists A \in \mathcal{C}(r) \text{ such that } \left| \|A_\Omega\|_2^2 - \|A\|_{L_2(\Pi)}^2 \right| > \frac{1}{2} \|A\|_{L_2(\Pi)}^2 + \mathcal{E} \right\}.$$

Note that \mathcal{B} contains the complement of the event that we are interested in.

In order to estimate the probability of \mathcal{B} we use a standard peeling argument. Let $\nu = \frac{\log(d)}{0.0006 \log(6/5) p}$ and $\alpha = \frac{6}{5}$. For $l \in \mathbb{N}$ set

$$S_l = \left\{ A \in \mathcal{C}(r) : \alpha^{l-1} \nu \leq \|A\|_{L_2(\Pi)}^2 \leq \alpha^l \nu \right\}.$$

If the event \mathcal{B} holds for some matrix $A \in \mathcal{C}(r)$, then A belongs to some S_l and

$$\begin{aligned} \left| \|A_\Omega\|_2^2 - \|A\|_{L_2(\Pi)}^2 \right| &> \frac{1}{2} \|A\|_{L_2(\Pi)}^2 + \mathcal{E} \\ &> \frac{1}{2} \alpha^{l-1} \nu + \mathcal{E} \\ &= \frac{5}{12} \alpha^l \nu + \mathcal{E}. \end{aligned} \quad (37)$$

For $T > \nu$ consider the following set of matrices

$$\mathcal{C}(r, T) = \left\{ A \in \mathcal{C}(r) : \|A\|_{L_2(\Pi)}^2 \leq T \right\}$$

and the following event

$$\mathcal{B}_l = \left\{ \exists A \in \mathcal{C}(r, \alpha^l \nu) : \left| \|A_\Omega\|_2^2 - \|A\|_{L_2(\Pi)}^2 \right| > \frac{5}{12} \alpha^l \nu + \mathcal{E} \right\}.$$

Note that $A \in S_l$ implies that $A \in \mathcal{C}(r, \alpha^l \nu)$. Then (37) implies that \mathcal{B}_l holds and we get $\mathcal{B} \subset \cup \mathcal{B}_l$. Thus, it is enough to estimate the probability of the simpler event \mathcal{B}_l and then apply the union bound. Such an estimation is given by the following lemma. Its proof is given in Appendix D. Let

$$Z_T = \sup_{A \in \mathcal{C}(r, T)} \left| \|A_\Omega\|_2^2 - \|A\|_{L_2(\Pi)}^2 \right|.$$

Lemma 10. *We have that*

$$\mathbb{P} \left(Z_T \geq \frac{5}{12} T + 44 p^{-1} \left[r (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right] \right) \leq 4e^{-c_1 p T}$$

with $c_1 \geq 0.0006$.

Lemma 10 implies that $\mathbb{P}(\mathcal{B}_l) \leq 4 \exp(-c_1 p \alpha^l \nu)$. Using the union bound we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &\leq \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\ &\leq 4 \sum_{l=1}^{\infty} \exp(-c_1 p \alpha^l \nu) \\ &\leq 4 \sum_{l=1}^{\infty} \exp(-c_1 p \nu \log(\alpha) l) \end{aligned}$$

where we used $e^x \geq x$. We finally compute for $\nu = \frac{\log(d)}{0.0006 p \log(6/5)}$

$$\mathbb{P}(\mathcal{B}) \leq \frac{4 \exp(-c_1 p \nu \log(\alpha))}{1 - \exp(-c_1 p \nu \log(\alpha))} = \frac{4 \exp(-\log(d))}{1 - \exp(-\log(d))}.$$

This completes the proof of Lemma 9.

D Proof of Lemma 10

We will start by showing that Z_T concentrates around its expectation and then we will upper bound the expectation. Recall that by definition,

$$Z_T = \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} \eta_{ij} A_{ij}^2 - \mathbb{E} \left(\sum_{(i,j)} \eta_{ij} A_{ij}^2 \right) \right|.$$

We use the following Talagrand's concentration inequality :

Theorem 11. Suppose that $f : [-1, 1]^N \rightarrow \mathbb{R}$ is a convex Lipschitz function with Lipschitz constant L . Let Ξ_1, \dots, Ξ_N be independent random variables taking value in $[-1, 1]$. Let $Z := f(\Xi_1, \dots, \Xi_N)$. Then for any $t \geq 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq 16L + t) \leq 4e^{-t^2/2L^2}.$$

For a proof see [22] and [8]. Let $f(x_{11}, \dots, x_{m_1 m_2}) := \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} (x_{ij} - p_{ij}) A_{ij}^2 \right|$.

It is easy to see that $f(x_{11}, \dots, x_{m_1 m_2})$ is a Lipschitz function with Lipschitz constant $L = \sqrt{p^{-1}T}$. Indeed,

$$\begin{aligned} & |f(x_{11}, \dots, x_{m_1 m_2}) - f(z_{11}, \dots, z_{m_1 m_2})| \\ &= \left| \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} (x_{ij} - p_{ij}) A_{ij}^2 \right| - \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} (z_{ij} - p_{ij}) A_{ij}^2 \right| \right| \\ &\leq \sup_{A \in \mathcal{C}(r, T)} \left\| \sum_{(i,j)} (x_{ij} - p_{ij}) A_{ij}^2 - \sum_{(i,j)} (z_{ij} - p_{ij}) A_{ij}^2 \right\| \\ &\leq \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} (x_{ij} - p_{ij}) A_{ij}^2 - \sum_{(i,j)} (z_{ij} - p_{ij}) A_{ij}^2 \right| \\ &\leq \sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i,j)} (x_{ij} - z_{ij}) A_{ij}^2 \right| \\ &\leq \sup_{A \in \mathcal{C}(r, T)} \sqrt{\sum_{(i,j)} \pi_{ij}^{-1} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} A_{ij}^4} \\ &\leq \sqrt{p^{-1}} \sup_{A \in \mathcal{C}(r, T)} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} A_{ij}^2} \\ &\leq \sqrt{p^{-1}T} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2} \end{aligned}$$

where we used $||a| - |b|| \leq |a - b|$, $\|A\|_\infty \leq 1$ and $\|A\|_{L_2(\Pi)}^2 \leq T$. Now, Theorem 11 and $2\sqrt{p^{-1}T} \leq T + p^{-1}$ imply

$$\mathbb{P}\left(Z_T \geq \mathbb{E}(Z_T) + 768p^{-1} + \frac{1}{12}T + t\right) \leq 4e^{-t^2 p/2T}.$$

Taking $t = \frac{1}{9}\left(\frac{1}{3}T\right)$ we get

$$\mathbb{P}\left(Z_T \geq \mathbb{E}(Z_T) + 768p^{-1} + \frac{1}{9}\left(\frac{5}{12}T\right)\right) \leq 4e^{-c_1 p T} \quad (38)$$

with $c_1 \geq 0.0006$.

Next we bound the expectation $\mathbb{E}(Z_T)$. Using a standard symmetrization argument (see e.g. [18]) we obtain

$$\begin{aligned}\mathbb{E}(Z_T) &= \mathbb{E} \left(\sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i, j)} \eta_{ij} A_{ij}^2 - \mathbb{E}(\eta_{ij} A_{ij}^2) \right| \right) \\ &\leq 2\mathbb{E} \left(\sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i, j)} \epsilon_{ij} \eta_{ij} A_{ij}^2 \right| \right)\end{aligned}$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence. Then, the contraction inequality (see e.g. [16, Theorem 2.2]) yields

$$\mathbb{E}(Z_T) \leq 8\mathbb{E} \left(\sup_{A \in \mathcal{C}(r, T)} \left| \sum_{(i, j)} \epsilon_{ij} \eta_{ij} A_{ij} \right| \right) = 8\mathbb{E} \left(\sup_{A \in \mathcal{C}(r, T)} |\langle \Sigma_R, A \rangle| \right)$$

where $\Sigma_R = \sum_{(i, j)} \epsilon_{ij} \eta_{ij} X_{ij}$. For $A \in \mathcal{C}(r, T)$ we have that

$$\begin{aligned}\|A\|_* &\leq \sqrt{r} \|A\|_2 \\ &\leq \sqrt{r p^{-1}} \|A\|_{L_2(\Pi)} \\ &\leq \sqrt{r p^{-1} T}\end{aligned}$$

where we have used (3). Then, by the duality between nuclear and operator norms, we compute

$$\mathbb{E}(Z_T) \leq 8\mathbb{E} \left(\sup_{\|A\|_* \leq \sqrt{r p^{-1} T}} |\langle \Sigma_R, A \rangle| \right) \leq 8\sqrt{r p^{-1} T} \mathbb{E}(\|\Sigma_R\|).$$

Finally, using

$$\frac{1}{9} \left(\frac{5}{12} T \right) + 8\sqrt{r p^{-1} T} \mathbb{E}(\|\Sigma_R\|) \leq \left(\frac{1}{9} + \frac{8}{9} \right) \frac{5}{12} T + 44 r p^{-1} (\mathbb{E}(\|\Sigma_R\|))^2$$

and the concentration bound (38) we obtain that

$$\mathbb{P} \left(Z_T \geq \frac{5}{12} T + 44 r p^{-1} \left[r (\mathbb{E}(\|\Sigma_R\|))^2 + 18 \right] \right) \leq 4e^{-c_1 p T}$$

with $c_1 \geq 0.0006$ as stated.

E Proof of Lemma 1

It is easy to see that

$$\|(M_{k+1} - M_k)_{\bar{\Omega}}\| \leq \|(M_{k+1} - M_k)_{\bar{\Omega}}\|_2 \leq \|M_{k+1} - M_k\|_2$$

and

$$\|M_{k+1} - M_k\|_\infty \leq \|M_{k+1} - M_k\|_2.$$

Thus, it is enough to show (8). The proof of (8) is close to the proof of Lemma 4 in [19].

Let us denote for by \tilde{M}_k the solutions produced by Algorithm 1 after soft-thresholding step and before truncating step (6). We have that

$$\|M^{k+1} - M^k\|_2 \leq \|\tilde{M}^{k+1} - \tilde{M}^k\|_2 \leq \|(M^k - M^{k-1})_{\bar{\Omega}}\|_2 \leq \|M^k - M^{k-1}\|_2 \quad (39)$$

where in the second inequality we used the following result (see, for example, Lemma 3 in [19])

Proposition 12. *The soft-thresholding operator $S_\lambda(\cdot)$ satisfies the following: for any W_1, W_2*

$$\|S_\lambda(W_1) - S_\lambda(W_2)\|_2 \leq \|W_1 - W_2\|_2.$$

The inequality (39) implies that the sequence $\{\|M^k - M^{k-1}\|_2\}_{k \geq 1}$ converges. It remains to show that it converges to zero. Note that the inequalities (39) imply that

$$\|M^k - M^{k-1}\|_2^2 - \|(M^{k+1} - M^k)_{\bar{\Omega}}\|_2^2 = \|(M^{k+1} - M^k)_\Omega\|_2^2 \rightarrow 0.$$

So, we only need to show that $\|(M^{k+1} - M^k)_{\bar{\Omega}}\|_2 \rightarrow 0$.

We put

$$Q(A, B) = \frac{1}{2}\|(Y - B)_\Omega\|_2^2 + \frac{1}{2}\|(A - B)_{\bar{\Omega}}\|_2^2 + \lambda\|B\|_*.$$

Note that (7) implies

$$\begin{aligned} Q(M^k, \tilde{M}^k) &\geq Q(M^k, \tilde{M}^{k+1}) \\ &= \frac{1}{2}\|(Y - \tilde{M}^{k+1})_\Omega\|_2^2 + \frac{1}{2}\|(M^k - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 + \lambda\|\tilde{M}^{k+1}\|_* \\ &\geq \frac{1}{2}\|(Y - \tilde{M}^{k+1})_\Omega\|_2^2 + \frac{1}{2}\|(M^{k+1} - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 + \lambda\|\tilde{M}^{k+1}\|_* \\ &= Q(M^{k+1}, \tilde{M}^{k+1}) \end{aligned} \quad (40)$$

where in the last inequality we used that

$$M_{ij}^{k+1} = \begin{cases} \tilde{M}_{ij}^{k+1} & \text{if } |\tilde{M}_{ij}^{k+1}| \leq a \\ a & \text{if } \tilde{M}_{ij}^{k+1} > a \\ -a & \text{if } \tilde{M}_{ij}^{k+1} < -a. \end{cases} \quad (41)$$

The inequality (40) shows that the sequence $\{Q(M^k, \tilde{M}^k)\}_{k \geq 1}$ converges. This and (40) yield

$$\begin{aligned} Q(M^k, \tilde{M}^{k+1}) - Q(M^{k+1}, \tilde{M}^{k+1}) \\ = \frac{1}{2} \|(M^k - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 - \frac{1}{2} \|(M^{k+1} - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 \rightarrow 0. \end{aligned} \quad (42)$$

Now, it is easy to see that

$$\|(M^k - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 - \|(M^{k+1} - \tilde{M}^{k+1})_{\bar{\Omega}}\|_2^2 \geq \|(M^k - M^{k+1})_{\bar{\Omega}}\|_2^2. \quad (43)$$

Indeed, for (i, j) in $\bar{\Omega}$ such that $M_{ij}^{k+1} = \tilde{M}_{ij}^{k+1}$ we have that

$$\left(M_{ij}^k - \tilde{M}_{ij}^{k+1}\right)^2 - \left(M_{ij}^{k+1} - \tilde{M}_{ij}^{k+1}\right)^2 = \left(M_{ij}^k - M_{ij}^{k+1}\right)^2$$

and for (i, j) in $\bar{\Omega}$ such that $M_{ij}^{k+1} \neq \tilde{M}_{ij}^{k+1}$ we have that

$$\left(M_{ij}^k - \tilde{M}_{ij}^{k+1}\right)^2 - \left(M_{ij}^{k+1} - \tilde{M}_{ij}^{k+1}\right)^2 \geq \left(M_{ij}^k - M_{ij}^{k+1}\right)^2$$

where we used (41). Now (42) together with (43) imply (8) which completes the proof of Lemma 1.

F Proof of Lemma 4

In order to prove (10), we use the following remarkable bound on the spectral norms of random matrices. It is obtained by extension to rectangular matrices via self-adjoint dilation of Corollary 3.12 and Remark 3.13 in [2] (cf., Section 3.1 in [2]).

Proposition 13 ([2]). *Let A be the $m_1 \times m_2$ rectangular matrix whose entries A_{ij} are independent centered bounded random variables. Then, for any $0 < \epsilon \leq 1/2$ there exists a universal constant c_ϵ such that, for every $t \geq 0$*

$$\mathbb{P} \left\{ \|A\| \geq (1 + \epsilon) 2\sqrt{2}(\sigma_1 \vee \sigma_2) + t \right\} \leq (m_1 \wedge m_2) \exp \left(\frac{-t^2}{c_\epsilon \sigma_*^2} \right)$$

where we have defined

$$\begin{aligned} \sigma_1 &= \max_i \sqrt{\sum_j \mathbb{E}[A_{ij}^2]} \\ \sigma_2 &= \max_j \sqrt{\sum_i \mathbb{E}[A_{ij}^2]} \\ \sigma_* &= \max_{ij} |A_{ij}|. \end{aligned}$$

We apply Proposition 13 to $\Sigma = \sum_{(i,j)} \eta_{ij} \xi_{ij} X_{ij}$. We compute

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}[\eta_{ij}^2 \xi_{ij}^2]} = \sigma \max_i \sqrt{\pi_i} \quad \text{and} \quad \sigma_2 = \sigma \max_j \sqrt{\pi_{\cdot j}}.$$

Bound (4) implies that $\sigma_1 \vee \sigma_2 \leq \sigma \sqrt{L}$. On the other hand, Assumption 1 implies $\max_{ij} |\eta_{ij} \xi_{ij}| \leq b$. Now, taking in Proposition 13 $\epsilon = 1/2$ we get (10).

In order to prove (11) we use the following result

Proposition 14 (Corollary 3.3 in [2]). *Let A be the $m_1 \times m_2$ rectangular matrix with A_{ij} independent centered bounded random variables. Then, there exists a universal constant C^* such that,*

$$\mathbb{E} \|A\| \leq C^* \left\{ \sigma_1 \vee \sigma_2 + \sigma_* \sqrt{\log(m_1 \wedge m_2)} \right\}$$

where $\sigma_1, \sigma_2, \sigma_*$ are defined in Proposition 13.

We apply Proposition 14 to $\Sigma_R = \sum_{(i,j)} \eta_{ij} \epsilon_{ij} X_{ij}$ where $\{\epsilon_{ij}\}$ is i.i.d. Rademacher sequence. We have that $\sigma_1 \vee \sigma_2 \leq \sqrt{L}$ and $\sigma_* \leq 1$, then Proposition 14 implies (11).

References

- [1] L. Balzano, R. Nowak, and B. Recht. Absolute and monotonic norms. *Numerische Mathematik*, 3:257–264, 1961.
- [2] A. S. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *ArXiv e-prints*, August 2014.
- [3] J. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [4] T. T. Cai and W. Zhou. Matrix completion via max-norm constrained optimization. *URL = <http://dx.doi.org/10.1007/978-1-4612-0537-1>*, 201E.
- [5] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of IEEE*, 98(6):925–936, 2009.
- [6] E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.
- [7] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- [8] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 02 2015.

- [9] C. Dhanjal, R. Gaudel, and S. Cl  men  on. Online matrix completion through nuclear norm regularisation. *SIAM International Conference on Data Mining*, pages 623–631, 2014.
- [10] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. *Journal 24nd Annual Conference on Learning Theory (COLT)*, 2011.
- [11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011.
- [12] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010.
- [13] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.
- [14] O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electron. J. Statist.*, 5:1161–1183, 2011.
- [15] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.
- [16] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008,   cole d’  t   de Probabilit  s de Saint-Flour. [Saint-Flour Probability Summer School].
- [17] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.
- [18] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
- [19] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.
- [20] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.
- [21] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.
- [22] M. Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 01 1996.

- [23] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
- [24] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.
- [25] Z. Wang, M. Lai, Z. Lu, W. Fan, H. Davulcu, and J. Ye. Orthogonal rank-one matrix pursuit for low rank matrix completion. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 91–99, 2014.
- [26] G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170(0):33 – 45, 1992.